2/PRTS.

Method and arrangement for translation of information

The invention relates to a method and an arrangement for translating information given as a character string in a first language into a character string in a second language. The invention is advantageously implemented in machine translation of text information.

There are previously known methods for linguistically based machine translation of text information. In these methods, the syntax of each language is exactly programmed, so that each language will require a program algorithm of its own. For the storage of vocabularies in different languages, a centralised high-capacity translation memory is used. The EuroTra translation system of the European Union can be mentioned as an example of such a method. Such previously known methods have a number of drawbacks. Exact syntax programming requires most extensive programming operations. Such a syntax algorithm, as well as the necessary translation memory, require a large memory space in the database. Since a translation method operating in this manner is complex, translating within a reasonable time requires an extremely high-powered computer. Due to these shortcomings, the equipment suitable for translation is expensive. Known methods also involve the drawback that updating of the translation algorithm requires programming and updating of the computer program each time.

The object of the present invention is to provide a solution for the translation of information which enables the prior art inconveniences described above to be overcome.

One idea of the invention is to divide the information to be translated into structural segments and to do the translation by structural segments. The translation is performed on the basis of model segments and rules stored in the knowledge base. The data contained in the knowledge base are advantageously increased so that, in the process of translating, whenever necessary, the user is asked to provide translations of new structural segments over a user interface, these translations being subsequently stored as model segments in the knowledge base. Owing to the solution provided by the invention, the translating equipment requires a smaller memory capacity and a lower processor speed. Moreover, far less programming is required and the operation of the equipment can be developed without program updating.

The method of the invention for machine translation of information given as a character string in a first language into a character string in a second language is characterised by

- storing model segments in the form of character strings in the first language in the knowledge base and, logically connected to these, model segments in the form of character strings in the second language,
- identifying a structural segment in the character string of said first language following a first rule,
- comparing said identified structural segment with model segments in the form of character strings in the first language stored according to a second rule,
- striving to select one model segment on the basis of said comparison,
- reading a model segment in the form of a character string in the second language logically connected to the selected model, i.e. equivalent segment and
- translating said structural segment into said translation segment in the form of a character string in the second language on the basis of said equivalent segment and a third rule.

The arrangement of the invention for translating information given as a character string in a first langauge into a character string in a second language is characterised in comprising

- knowledge base means for storing model segments in the form of said character strings in the first language and, in logical connection with these, for storing equivalent segments in the form of character strings in the second language, and for storing a first, second and third rule,
- means for identifying structural segments in said information given as a character string in the first language following a first rule,
- means for comparing said identified structural segment with the stored model segments in the form of character strings in the first language following a second rule,
- means for selecting one model segment on the basis of said comparison,
- means for reading a model segment, i.e. equivalent segment, in the form of a character string in the second language, logically connected to the selected model, in said knowledge base means and
- means for translating said structural segment into said translation segment in the form of a character string in the second language on the basis of said equivalent segment and the third rule, said translation segment representing the information to be given in said second language.

Preferred embodiments of the invention are described in the dependent claims.

The invention is described in greater detail below with the aid of the accompanying drawings, of which

figure 1 is a flow chart of a method in accordance with the invention for translating information,

figure 2 is a block diagram of an arrangement in accordance with the invention for translating information,

figure 3 illustrates text information divided into structural segments,

figure 4 illustrates the translating process of one structural segment with a close model segment appearing in the knowledge base and

figure 5 illustrates the translating process of a structural segment with no close model segment appearing in the knowledge base.

Figure 1 illustrates a method in accordance with the invention for translating information. First the information to be translated is read, block 101, and is divided into structural segments according to a first rule, block 102. Subsequently, the first structural segment is read in the untranslated information, block 103. The read structural segment is compared with the model segments stored in the knowledge base, blocks 104 and 110. The comparison is then performed according to a second rule, which determines whether the model segment is close to the structural segment to be translated. If a model segment closely related to this particular structural segment is found in the knowledge base, a model segment i.e. equivalent segment, in the second language logically connected to the close model, block 121, is read in the knowledge base. After this, a translation segment translated into the second language is formed from the structural segment to be translated on the basis of the read equivalent segment following a third rule, block 122. After this it is checked whether there are still untranslated structural segments, block 123. If there are still untranslated structural segments, the process returns to block 103, where the following untranslated structural segment is read for translation. If there are no untranslated structural segments left in block 123, the translation segments are arranged into sentences according to a fourth rule, and the translated information is then stored. The stored information can be further displayed, e.g. on a screen, or printed out e.g. on paper or a disc, block 124.

If no model segment close to the structural segment is found in the knowledge base in block 110, this particular structural segment is displayed over a user interface means, i.e. a display screen, block 131. The user then feeds the translation of the structural segment, i.e. the equivalent segment, block 132. The structural segment and the equivalent segment are stored for future use as model segments in the knowledge base, blocks 133, 134. After this the process proceeds to block 123 to continue as explained above. In this case, the equivalent segment is usually directly a translation segment, if the user has been asked to give the translation of the structural segment in the form of the original information. Thus the operation of block 122 is not indispensable in this case.

Said first rule, by which the structural segments are identified, can be based for instance on the identification of ”intermediate words” or cases. Intermediate words are for instance prepositions and particles, which usually form standard character strings. Thus, they can be identified by simply comparing the character strings forming each word e.g. with the above known character strings forming an intermediate word. The identification of cases can be performed e.g. with the aid of suffixes by comparing the last characters of the words with known suffixes. As well known, the character strings forming a word can be separated by means of punctuation. Since a structural segment may advantageously comprise several words, it may also include one or more punctuation marks.

In its most straightforward version, said second rule, by which a structural segment is compared with the model segments, may imply similarity. In this case, exactly the same model segment as the present structural segment to be translated is searched in the knowledge base. However, considering the memory space required for the knowledge base, is is preferable not to store the different cases of e.g. the model segment separately in the knowledge base, but to identify also a model segment having a different case following the second rule. In this situation, the equivalent segment logically connected to the model segment should also be put in the case needed in order to generate a translation segment. This is done according to the third rule, which consequently covers information about the cases of the language in question.

In many cases, said fourth rule, by which the translation segments are arranged in translated sentences, implies placing the translation segments into the same order in which the structural segments to be translated were in the first language. Yet this order may depend on the language, and hence also said fourth rule is language-specific.

In the storage of the model segments, a type identifier of the model segment can also be advantageously stored. In this case, the type identifier is stored in logical connection with each model segment. If type identifiers are used, various rules can be applied the identification and translation of the structural segment on the basis of the model segment, depending on the type of the structural segment. Types of structural segments are e.g. the object of an action, a proper name, a verb, a place word, an adjective or an idiom. If type identifiers are used, the user is also asked to indicate the type to which the particular structural segment and its translation pertain as the structural segment is translated.

One idea of the invention is to update the knowledge base in the interactively operated translation process. It should be noted that the updating of the knowledge base is not necessarily confined to the storage of new model or equivalent segments, but the rules mentioned above can also be advantageously updated. The updating is then performed e.g. in connection with the translation of a new structural segment fed by the user by identifying the regularity of the input translation.

The translation of one piece of information from a first language into a second language has been described above. The previous updatings of the knowledge base are advantageously utilised in the translation of the subsequent pieces of information. Thus, the process of the invention for translating successive first and second pieces of information may comprise e.g. the following steps:

- reading first information given as a character string in the first language,
- performing the translation of the first information given as a character string in said first language on the basis of data in the knowledge base into first information given as a character string in the second language to the extent this is feasible in terms of the data available in the knowledge base,
- determining the additional data required to complete the translation of the first information given as a character string in the first language into first information given as a character string in the second language,
- feeding said additional data in the knowledge base with a view to update the knowledge base,

- finishing the translation of the first information given as a character string in the first language into first information given as a character string in the second language,
- storing said first information given as a character string in the second language,
- storing the second information given as a character string in the first language,
- performing the translation of the second information given as a character string in said first language on the basis of said updated data in the knowledge base into second information given as a character string in the second language.

Figure 2 is a block diagram of a device arrangement of the invention for the translation of information. The arrangement comprises a disc station 21, a display screen 22 and a keyboard 23 as interface means connected to processor 20. By means of the disc station, information to be translated can be fed from the disc to the device and the translated information can be stored on the disc for use in other devices. The information in question can be transferred between the device and other data processing equipment also over a bus I/O. Display screen 22 can be used to display such structural segments to the user for which no translation is found in the knowledge base. The user can feed the translation of such a structural segment by using keyboard 23. The interface means mentioned above can also be used in the revision and correction of translated information.

The device shown in figure 2 also comprises an electric memory 24 for temporary storage of structural segments and translation segments, among other things. In addition, the device comprises a mass storage 25 for the storage of the knowledge base, i.e. model segments, type identifiers and rules, as well as programs. For instance a hard disc drive or an optical disc drive can be used as a mass storage. The components mentioned above can be provided by making previously known computer components operate in accordance with the invention using special software. Character strings and other data are advantageously transferred as electric signals between the components.

The implementation of the invention is by no means confined to the components described above, by contrast, the arrangement of the invention can have many different configurations, which this description enables a person skilled in the art to design.

Figure 3 illustrates an English sentence divided into structural segments 31, 32, 33 and 34. As shown in the figure, a structural segment typically comprises successive closely related words in a sentence. Thus a structural segment often includes a punctuation mark separating the words as well.

Figure 4 illustrates the translation of the first structural segment of the sentence appearing in figure 3 with the aid of one solution of the invention. In the figure, the structural segment 42 to be translated is stored in translation memory 41 and this structural segment is compared with the model segments stored in knowledge base 44. In the case illustrated in figure 4, this particular structural segment has been previously stored in the knowledge base as model segment 45, which is found in the comparison. If, for instance, the present information is to be translated into Finnish, the Finnish model segment 46 logically connected to the English model segment mentioned above is read in the knowledge base. In figure 4 the double line connecting model segments 45 and 46 illustrates a logical connection. When the Finnish model segment has been read it is stored as a translation segment in the translation memory.

Figure 5 illustrates the translation of the second structural segment shown in figure 3 with the aid of a solution of the invention. In this case, no English structural segment to be translated nor any Finnish equivalent segment has been previously stored as a model segment in the knowledge base. In this case, structural segment 52 to be translated, stored in translation memory 51, is compared with the model segments in the knowledge base, and if the desired equivalent segment is not found in the knowledge base, the structural segment 58 to be translated is shown on the display screen of interface 57. After this, the user feeds the translation 59 of structural segment 58 over the interface in knowledge base 54. In this manner, an English and a Finnish model segment are stored in logical connection in the knowledge base. Then the Finnish translation of the structural segment is stored as a translation segment 53 in translation memory 51.

Should the structural segments mentioned above reappear in the input information, corresponding model and equivalent segments will be found in the knowledge base, and there will be no need to ask the user to repeat them. If, however, the following input information contains the sentence ”we have expanded our operation largely in Finland”, ”largely” would be a new structural segment. If no close model segment has been previously stored in the knowledge base, the user is asked to give the translation of it and ”largely” is stored as a model segment in the knowledge base, and in logical connection with this, also the input translation fed by the user.

It should be noted that the operation of the equipment can be arranged so that the translation process is first performed by machine for the entire information to the extent allowed by the model segments stored in the knowledge base. After this the user can feed the necessary translations of new structural segments in the knowledge base. Such an arrangement has the advantage of the user not having to stay by the computer waiting for the translation process to be completed, but he/she may update the knowledge base with one single input at any suitable moment.

The model segments can be stored in the knowledge base as pairs of segments, specific pairs of model segments being stored for each language pair. Another way of proceeding is to logically connect model segments in several languages, so that the same model segments can be used as such in the translation of several language pairs. In this case, the model segments of each language can be fed as an input in the knowledge base each time they appear for the first time in the language in question. When input information is then fed in the knowledge base during the translation of one language pair, the information contained in the knowledge base will automatically increase also in the other language pairs.

The solution of the invention is not language-specific on principle, but can be applied to any language pair. Nor is the implementation of the invention restricted to "natural" languages used in ordinary communication, since it can be used to translate any language consisting of character strings into a second language consisting of character strings. Programming languages and data exchange protocols may be mentioned as examples of such other languages.

The solution of the invention has many advantages over prior art. Its operation requires but little language-specific knowledge for the division of the language into structural segments. A second advantage of the solution consists in additional information being collected in the memory during the process, so that the device "learns" new pairs of model segments and rules. Thus, with a straightforward configuration and a small amount of programming and updating it is possible to provide an efficient means for machine translation.

The solution of the invention is well adapted for use in situations where the arrangement of the invention is used to meet the needs of several users. In this case, the arrangement preferably comprises several interfaces, which may communicate with the knowledge base e.g. over a data transmission network. The knowledge base can then preferably be decentralised in such a way that the first, i.e. the main knowledge base can be used by a given user group, and the second, i.e. subknowledge base, may be used only by a few in such a given user group. This enables different users to update their own knowledge base e.g. with special vocabularies or expressions, without such knowledge suitable for special purposes being used by other users.

In such a decentralised knowledge base, the updating of the first, i.e. the main knowledge base can be performed from the second, i.e. subknowledge bases. Data stored in the second knowledge bases are then transferred to the first knowledge base by predetermined criteria. One such criterion may be the incidence of specific data. The data exchange between the knowledge bases can also take place with one common main knowledge database administrator checking and approving each data to be transferred.

A number of embodiments of the solution in accordance with the invention has been described above. The principle of the invention can, of course, be varied within the scope of protection of the claims, for instance regarding details of the embodiment and fields of application.